一种基于相对相似性提高推荐总体多样性的协同过滤算法

姜书浩 [1, 2]　张立毅 [1, 2]　张志鑫 [2]

[1](天津大学电子信息工程学院　天津　300072)

[2](天津商业大学信息工程学院　天津　300134)

摘要:【目的】以提高推荐系统的总体多样性为出发点，解决因为用户评分数据分布不均和稀疏造成的误差从而影响推荐精确性和多样性问题。【方法】根据用户间共同评分项目的数量，通过加权计算得出相对相似性指数，修正相似性计算方法，进而优化预测评分算法，在保证推荐精确性的前提下提高总体多样性，提升企业的长尾营销效果。【结果】实验结果表明，当评分阈值为 3.5，最近邻数目为 20 时，本文方法在 MovieLens 数据集上的计算结果相对于采用传统的余弦相似性计算结果，总体多样性提高了 114，精确性提高了 6.5%。【局限】仅适用于基于最近邻的协同过滤算法，并不涉及其他推荐技术。【结论】该方法有效地提高了推荐的总体多样性，获得推荐精确性和总体多样性用户相对满意度都较高的推荐结果。

关键词: 总体多样性　相对相似性　协同过滤

分类号: TP301

1　引　言

对于每一个用户来说，如何从海量信息中甄别出有用的信息，十分困难但却又非常重要。个性化推荐系统是解决这一问题的重要方法，它帮助用户从庞大的数据集合中选择最合适的信息。推荐系统通过确定用户的偏好[1]，向特定用户推荐最适合他的或者他最感兴趣的项目。推荐算法主要包括基于内容的推荐、协同过滤推荐以及混合推荐等。推荐系统广泛应用于电影、音乐、图书、旅游、电子商务[2]、社交[3]和网络搜索等方面。精确性是评价推荐系统的一个重要指标，它是评价推荐给用户的项目是否是最适合的，但是这种方式推荐的信息用户很可能早已从其他渠道得到，因此很多情况下没有必要。而评价推荐系统优劣的另一个重要指标多样性，越来越受到研究人员和用户的关注，多样性反映的是推荐项目种类的差异性。一些学者甚至称，多样性在某些情况下超过精确性能给用户带来更好的满意度。

推荐系统的精确性和多样性是完全不同的两个方面，一个好的推荐系统应该兼顾这两个推荐标准，然而这两个标准是相互制约的。如果大幅提高推荐结果的多样性，精确性必然会受到影响，从而使推荐的结果相关性不强；如果推荐结果的精确性较高，多样性同样受损，推荐结果会出现较高的相似性而显得呆板。已经有非常多的研究关注如何提高推荐系统的多样性，但是这些研究中更多专注于提高某特定用户推荐列表的多样性，称为个体多样性，除此之外，多样性还有一个指标称之为总体多样性，总体多样性可以被理解为向不同的用户推荐不同项目的数量[1, 4]。有效提高推荐结果的总体多样性，不仅能满足用户的个性化体验要求，而且可以有效提升长尾营销的效果[5]，帮助企业实现利润最大化。总体多样性是不直接关系到个体多样性的。

协同过滤推荐系统中依据用户的评分数据做出推

通讯作者：姜书浩，ORCID：0000-0002-7706-063X，E-mail：mr_jiang1980@163.com。

荐，而评分数据集经常存在数量分布不均以及稀疏性等问题，造成推荐结果产生误差，这种误差对于推荐的精确性和多样性都造成很大的影响，目前很多研究都是围绕如何提高推荐的精确性而展开，很少涉及多样性，尤其是总体多样性。本文主要探讨的是总体多样性，在确保推荐精确性的前提下，尽力提高系统的推荐总体多样性。

2　文献综述

目前国内外的研究将多样性的定义分为两类：个体多样性与总体多样性。个体多样性是从单个用户的角度而言的度量标准，目标是对于特定用户，尽量推荐一些彼此相似度很低但又符合该用户兴趣的商品。针对多样性的推荐目前已经成为非常热门的研究领域，研究人员提出了各种提高推荐多样性的方法，但是基本上都是以牺牲一定的精确性为代价，而且多数的研究都集中在个体多样性[6-9]。

总体多样性反映的是推荐系统向不同用户推荐不同种类商品的能力[4, 10]。不同于个体多样性，总体多样性的评价需要对所有用户进行。虽然之前文中也提到过总体多样性与个体多样性没有直接的关系，但是总体多样性确实是一个更加宽泛的概念。也有一些研究是针对总体多样性进行的，Lacerda 等提出一种基于用户兴趣建模，从推荐精确性、新颖性和多样性等角度设计推荐系统，提高了推荐的总体多样性，向用户推荐了一些评分次数较少的长尾商品[11]。Park 提出一种基于已知评分值或评分次数进行聚类的推荐方法，提高了一些长尾商品的预测评分，进而提高推荐总体多样性[12]。Adomavicius 等提出了一种改进的项目排序技术，以提高系统的总体多样性[4]。项目受欢迎度排名、反向预测评分值、邻域评级方差等一些排名方法为系统设计人员提供了更好的灵活性，并且可以与不同的评级预测算法结合，以获得更好的总体多样性[6]。Fleder 等研究了推荐系统对销售多样性的影响，研究结果表明，即使是一些知名的推荐系统也可能会导致销售多样性的减少，因为这些系统都是在销售和评分基础上推荐产品[13]。Bobadilla 等提出基于优化的方法，以提高总体多样性，包括贪婪算法、基于最大流的方法和整数规划方法[10]。贪婪算法是一个迭代过程，是将已经推荐的项目替换为高于阈值的项目。基于最大流的方法是一个基于图的算法，通过制定用户和项目之间的最大流问题以改善推荐多样性。整数规划方法是用精确性和多样性解决多准则优化问题。上述几种方法在总体多样性上虽然有所改善，但是精确性同时受到了较大的影响。王森提出一种提高系统的总体多样性和长尾商品的推荐率的推荐方法，综合考虑了商品预测值、商品流行度、商品的偏爱度等多个标准[14]。本文提出一种解决用户评分数据分布不均和存在稀疏性的情况下在保证精确性的前提下，提高推荐总体多样性的方法。

3　基于相对相似性的协同过滤推荐

本文采用的推荐方法是基于用户的协同过滤推荐算法，相似性计算是协同过滤推荐中的关键步骤，其计算的结果对 K 个最近邻的产生具有决定性作用，进而影响到预测评分。常使用的相似性计算方法包括余弦相似性、Pearson 相关系数、Jaccard 相似性等，本文使用余弦相似性计算用户 u 与其他用户之间的相似性，对相似性的优化问题也是针对余弦相似性而言。

3.1　余弦相似性及预测评分

多数协同过滤推荐系统都是采用余弦相似性，并且事实证明该算法在很多研究中的应用都非常成功。假设 U 为推荐系统的用户集，I 为要推荐给用户的项目集，R(u, i)为用户 u 对项目 i 的实际评分，R*(u, i)为用户 u 对项目 i 的预测评分，如公式(1)所示[7]:

$$\mathrm{Sim}(u,u')=\frac{\sum_{i\in I(u,u')} R(u,i)\times R(u',i)}{\sqrt{\sum_{i\in I(u,u')} R(u,i)^2}\times\sqrt{\sum_{i\in I(u,u')} R(u',i)^2}} \qquad (1)$$

其中，I(u, u’)表示用户 u 和 u’都已经评过分的项目集，经过相似性计算后，可获得最近邻居集 S(u)，R(u)为用户 u 的平均评分，则预测评分 R*(u, i)的计算如公式(2)所示[7]:

$$R^*(u,i)=\overline{R(u)}+\frac{\sum_{u'\in S(u)} \mathrm{Sim}(u,u')\times(R(u',i)-\overline{R(u')})}{\sum_{u'\in S(u)} |\mathrm{Sim}(u,u')|} \qquad (2)$$

研究发现在很多情况下采用余弦相似性提高推荐精确性时，其实无形中降低了推荐的总体多样性，这些情况正是余弦相似度面对用户评分数据不均以及数据稀疏时计算所存在的不足之处。

例 1　假设有 4 个用户 u1, u2, u3, u4 以及他们评价

的项目，L_n (u1) = {i1, i3, i5}，L_n (u2) = {i3, i5, i7}，L_n(u3)={i3}，L_n(u4)={i2, i3, i7}。

很显然，u3 是其他三个用户的最近邻，因为在这种情况下，他们仅有一个共同评分的项目 i3，因此，用户 u3 与其他用户的余弦相似性值为 1。因为通常认为如果两个用户有一个共同评分的项目，那么无论两个个体评分差异有多大，都认为他们的余弦相似性为 1。

例 2 假设有 4 个用户，其评分分别是 u1= {2, 2, 2}，u2 = {3, 3, 3}，u3 = {5, 5, 5}，u4= {2, 5, 3}。

那么用户 u4 与其他三个用户的相似性是相同的，在这种情况下其结果都是 0.9366。

当评分的数目非常有限的情况下，这类问题会更加严重，这是因为当有效评分不足以支持相似性计算时，运算产生的误差概率会明显增大。如果这种情况出现在用户 u 的最近邻，则用户 u 的预测评分 R*(u, i)会与平均评分相同，因为在这种情况下 R(u', i)与 $\overline{R(u')}$ 相同(见公式(2))。所以，用户的评分越少，预测评分出现误差就会越大，进而影响到推荐系统的精确性和总体多样性。

3.2 优化相似性算法

解决上述问题的方法是将评分少的用户(无论是评分总数少还是个体项目评分少)进行弱化，强化评分数据多的用户，具体的方法是考虑两个用户共同评分的数目，现实中两个用户共同评分的项目越多，说明其相似性相对越高，所以在进行相似性计算时应该将用户间的共同评分项目数目作为一个重要因素考量，基于这一原理设计相对相似性算法(Relative Similarity , RS)如下:

```
CR: 共同评分项目数, MCR: 最大共同评分项目数
输入: 用户集 U, 项目集 I, 任意两用户间的相似性 sim_user, any_user
输出: 相对相似性 RS
①CR=0, MCR=0
②for   user = 1 to |U|-1 do
③   MCR=0
④   for any_user=user+1 to |U| do
⑤      CR=|I(user, any_user)|
⑥      if   CR > MCR   then
⑦         MCR=CR
⑧      end
⑨   end
⑩end
⑪for user=1 to |U| do
⑫   for any_user=user+1 to   |U|   do
⑬      CR=|I(user, any_user)|
⑭      W=(CR/MCR)
⑮      RS=W*sim_user, any_user
⑯   end
⑰ end
```

该算法是在相似性计算后，对用户之间的相似性计算结果进行修正，它有三个输入参数: 用户集、项目集以及计算后的用户之间的相似性。算法的第①行首先定义两个变量，分别是当前用户与任意用户的共同评分项数目 CR 和当前用户与所有用户中共同评分项数目最大值 MCR; 算法的第④行到第⑨行的内循环是计算当前用户与其他用户的共同评分项目数的最大值，而第②行到第⑩行的运算结果是得出所有用户与其他用户的共同评分项目数的最大值，为算法的下半部分权值的使用做数据准备。算法的第二部分(第⑪-⑰行)是对之前计算的相似性结果进行加权修正，其中第⑭行设定任意用户 any_user 与当前用户 user 共同评分项目数量相对于当前用户 user 与所有用户共同评分项数目的最大值的比值作为这两个用户的计算权值 W，第⑮行使用该权值修正相似性计算结果，得出相对相似性 RS，当两个用户的 CR 相对较大时，W 的值趋向于 1，相似性较高，而 CR 较小时，W 值趋向于 0，相似性较低，符合之前的算法设想。按照修正后的相似性值计算的最近邻更加准确。如在 3.1 节的例 1 中 L_n (u1) = {i1, i3, i5}，L_n (u2) = {i3, i5, i7}，L_n(u3)={i3}，L_n(u4)={i2, i3, i7}，假设用户集中仅有这 4 个用户，则用户 u1 与用户集中其他用户最多的共同评分项目数为 2，所以用户 u1 与 u3 的相对相似性 RS 为 1/2=0.5，如果用户的数目更多，则 u1 与 u3 的相似性可能更低; 例 2 中，4 个用户共同为三个项目评分，所以 u4 与其他三个项目的相对相似性 RS 应为 0.9366/3=0.3122。对比前后两个数据，优化后的相对相似性更能真实地反映用户间的相似性。

在预测评分时使用相对相似性 RS，在上述问题的情况下，产生误导性的相似性会通过加权共同评分项目进行调整，使得具有更多共同评分的项目对相似性的计算权值影响更大，而较少共同评分的项目权值较小，这样具有误导性的相似性将不会被作为最近邻来考虑。经过修正后的预测评分计算如公式(3)所示:

$$R^*(u,i)=\overline{R(u)}+\frac{\sum_{u'\in S(u)} RS_{u,u'}*(R(u',i)-\overline{R(u')})}{\sum_{u'\in S(u)}|RS_{u,u'}|} \qquad (3)$$

采用该公式进行预测评分的计算，可以获得精确性和总体多样性相对用户满意度较高的推荐结果。

4 实验结果及评价

4.1 数据集

实验采用的数据集是公开的数据集 MovieLens① 的子集。MovieLens 数据集包含 943 个用户对 1 682 部电影的 100 000 个评分数据，评分范围为从 1 到 5。将数据子集划分为 80%的训练集和 20%的测试集。在两个数据集中分别实现下面的操作，创建用户项目矩阵，采用修改后的相似性方程计算协同过滤中的最近邻，最近邻确定后，进行预测评分。之后依据准则，即用户 u 对项目 i 的预测评分是否大于评分阈值，最终确定推荐的项目。

4.2 评价指标

本文算法设计目的是在保证精确性的前提下提高总体多样性，因此进行算法评价时应同时考虑精确性和多样性两个指标。对目标项目进行评分预测后，算法通过设定评分阈值生成最终推荐列表，相关联的阈值定义为 Tr，对于每一个预测评分，如果 R*(u, i)≥Tr，计算其推荐精确性如下[4]:

$$\text{accuracy}=\frac{\sum_{u\in U}|\text{result}(Ln(u))|}{\sum_{u\in U}|(L_n(u))|} \tag{4}$$

其中，$L_n(u)$={i1, i2, …, in}为推荐列表的前 n 个推荐项目，并且 result($L_n(u)$) = {i∈$L_n(u)$|R(u, i)≥Tr}为推荐项目中评分超过规定阈值的项目。

即使是准确率比较高的推荐系统也不能保证用户对其推荐结果满意，推荐系统中另一个需要重点关注的内容是推荐商品的种类，相关的评价指标是推荐系统的多样性。但是不同的研究人员评价多样性的指标各不相同，多样性分为个体多样性和总体多样性，个体多样性是用户内的多样性，是衡量推荐系统对一个用户推荐商品的多样性，总体多样性衡量推荐系统对不同用户推荐不同商品的能力。本文采用的总体多样性的计算公式如下[4]:

$$\text{diversity}=|U_{u\in U}L_n(u)| \tag{5}$$

①http://www.movielens.org/.

4.3 实验结果分析

实验结果分别对比了 MovieLens 数据集使用相对相似性计算前后推荐结果的精确性和总体多样性的变化情况。图 1 中三条虚线显示采用传统的相似性计算最近邻数分别为 10、20、50 的情况下精确性随评分阈值的变化情况。三条实线显示在同样的情况下采用相对相似性计算精确性的变化情况。

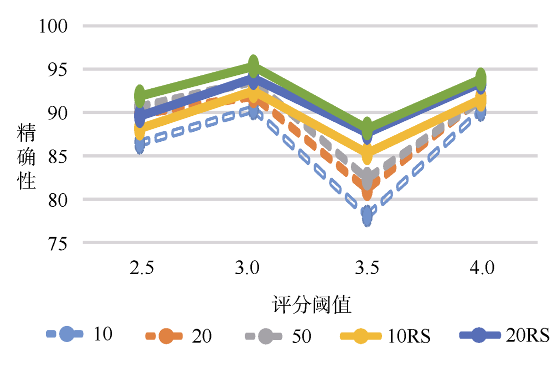


图 1 相对相似性使用前后精确性对比

图 1 中数据显示，在评分阈值相同的情况下，推荐的精确性随着最近邻数目的增加而提高，这一点两组数据类似。这说明最近邻的数目对于推荐精确性有正向的影响，但是最近邻的增多同时增加了预测评分的计算复杂度。虚实线的对比显示，采用相对相似性的推荐结果相对于未采用的推荐精度非但没有降低，反而有所提高，推荐精度在评分阈值为 3.5 时，提升效果明显。实验结果显示实现了算法设计时提出的保证推荐准确度的要求。

图 2 中同样用三条虚线和三条实线表示在采用相对相似性计算总体多样性的前后对比情况。可以看出，使用相对相似性的推荐结果总体多样性有了明显提高。实验数据显示评分阈值为 3.5 时，最近邻为 10 时，多样性由 98 提升到 221；最近邻为 20 时，多样性由 87 提高到 201；最近邻为 50 时，多样性由 79 提高到 127。

另外，图 2 中实线数据显示，当阈值较小时，总体多样性值相对于未使用相对相似性时有明显提升，当阈值逐渐增大时，尤其是达到 4 时，总体多样性值几乎接近于未使用之前的结果。

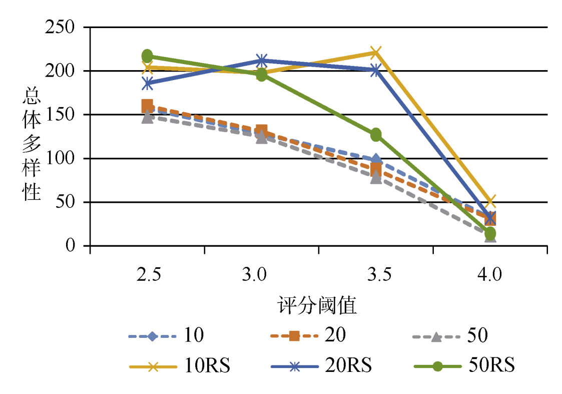


图 2 相对相似性使用前后总体多样性对比

从图 1 和图 2 中还可以看到，当阈值为 4 时，精确性虽然相对较高，但是多样性值大幅降低，可以得出结论：项目中评分阈值较高时，会导致推荐结果的种类相对比较集中。从实线数据的对比还可以看出，当阈值为 3.5 时，是推荐结果精确性最低而多样性最高的情况，从而也说明了两者相互制约的关系。

5 结 语

本文提出一种有效提高推荐总体多样性的方法，同时提出一种基于相对相似性分析推荐多样性的方式。从实验结果可以看出，本文提出的推荐模型相对于之前的研究方法，总体多样性与精确性随着阈值的增加都得到了有效的优化提升。同时，在阈值相对较高的情况(如 3、3.5)下，系统的推荐结果既保持了较高的精确性，同时多样性也得到较好的优化。

本文提出的方法主要出发点在于提高推荐系统的总体多样性，对于个体多样性的优化之前也做过相关研究，两种多样性的相关性以及综合提升方法是进一步研究工作的方向。

参考文献:

[1] Adomavicius G, Kwon Y. Optimization-based Approaches for Maximizing Aggregate Recommendation Diversity [J]. Informs Journal on Computing, 2014, 26(2): 351-369.

[2] Shambour Q, Lu J. An Effective Recommender System by Unifying User and Item Trust Information for B2B Applications [J]. Journal of Computer and System Sciences, 2015, 81(7): 1110-1126.

[3] Yigit M, Bilgin B E, Karahoca A. Extended Topology Based Recommendation System for Unidirectional Social Networks [J]. Expert Systems with Applications, 2015, 42(7): 3653-3661.

[4] Adomavicius G, Kwon Y. Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques [J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(5): 896-911.

[5] Núñez-Valdez E R, Lovelle J M C, Martínez O S, et al. Implicit Feedback Techniques on Recommender Systems Applied to Electronic Books [J]. Computers in Human Behavior, 2012, 28 (4) 1186-1193.

[6] Bradley K, Smyth B. Improving Recommendation Diversity [C]. In: Proceedings of the 12th Irish Conference on Artificial Intelligence and Cognitive Science. Maynooth, Ireland.2001.

[7] Zhang M, Hurley N. Avoiding Monotony: Improving the Diversity of Recommendation Lists [C]. In: Proceedings of the 2nd ACM Conference on Recommender Systems. ACM, 2008.

[8] Chen J, Liu Y, Hu J, et al. A Novel Framework for Improving Recommender Diversity [A]. // Behavior and Social Computing [M]. Springer International Publishing. 2013.

[9] Aytekin T, Karakaya M Ö. Clustering-based Diversity Improvement in Top-N Recommendation [J]. Journal of Intelligent Information Systems, 2014, 42(1): 1-18.

[10] Bobadilla J, Ortega F, Hernando A, et al. Recommender Systems Survey [J]. Knowledge Based Systems, 2013, 46: 109-132.

[11] Lacerda A, Ziciani N. Building User Profile to Improve User Experience in Recommender Systems [C]. In: Proceedings of the 6th ACM International Conference on Web Search and Data Mining. 2013.

[12] Park Y J. The Adaptive Clustering Method for the Long Tail Problem of Recommender Systems [J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(8): 1904-1915.

[13] Fleder D, Hosanagar K. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity [J]. Management Science, 2009, 55(5): 697-712.

[14] 王森.一种基于整体多样性增强的推荐算法[J]. 计算机工程与科学, 2006, 38(1): 183-187. (Wang Sen. A Recommendation Algorithm Based on Aggregate Diversity Enhancement [J]. Computer Engineering & Science, 2016, 38(1): 183-187.)

作者贡献声明:

姜书浩，张立毅：提出研究思路，设计研究方案；
姜书浩，张志鑫：进行实验；
张志鑫：采集、清洗和分析数据；

姜书浩: 起草论文;
张立毅, 姜书浩: 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: mr_jiang1980@163.com。
[1] 姜书浩. u.user.txt. MovieLens 用户原始数据.
[2] 姜书浩. u.item.txt. MovieLens 项目原始数据.
[3] 姜书浩. u.data.txt. MovieLens 用户评分原始数据.
[4] 姜书浩. u.genre.txt. MovieLens 电影类型原始数据.
[5] 姜书浩. 测试集推荐结果(使用传统余弦相似性结果).txt. 未使用 RS 测试集推荐结果.
[6] 姜书浩. 测试集推荐结果(使用相对相似性结果).txt. 使用 RS 测试集推荐结果.

收稿日期: 2016-08-15
收修改稿日期: 2016-09-19

New Collaborative Filtering Algorithm Based on Relative Similarity

Jiang Shuhao [1,2] Zhang Liyi [1,2] Zhang Zhixin [2]
[1](School of Electronic Information Engineering, Tianjin University, Tianjin 300072, China)
[2](Information Engineering College, Tianjin University of Commerce, Tianjin 300134, China)

Abstract: [**Objective**] The purpose of this study is to improve the overall diversity of the recommendation results. The proposed algorithm reduces errors caused by the uneven distribution and sparsity of user rating data, and then improves the recommendation accuracy and diversity. [**Methods**] We first generated the relative similarity index based on the number of common ratings and individual weights. Second, we modified the similarity calculation method, and the rating prediction algorithm. The proposed model improved the aggregated diversity and maintained the recommendation accuracy, which improved the marketing effects. [**Results**] The aggregated diversity index increased 114, the accuracy improved 6.5% on the MovieLens data compared with results generated by the traditional cosine similarity calculation, (the rating threshold was 3.5 and number of KNN is 20). [**Limitations**] This method was only applicable to collaborative filtering based on the nearest neighbor, and it did not include other recommendation techniques. [**Conclusions**] The proposed method effectively improves the diversity and accuracy of recommendation results, which significantly improves the user experience.

Keywords: Aggregate diversity Relative similarity Collaborative filtering